

# Sampling Online Social Networks via Heterogeneous Statistics

Xin Wang<sup>†</sup>, Richard T. B. Ma<sup>\*</sup>, Yinlong Xu<sup>†</sup>, Zhipeng Li<sup>†</sup>

<sup>†</sup> School of Computer Science and Technology, University of Science and Technology of China

<sup>\*</sup> School of Computing, National University of Singapore

{yixinxa, lizhip}@mail.ustc.edu.cn, tbma@comp.nus.edu.sg, ylxu@ustc.edu.cn

**Abstract**—Most sampling techniques for online social networks (OSNs) are based on a particular sampling method on a single graph, which is referred to as a statistic. However, various realizing methods on different graphs could possibly be used in the same OSN, and they may lead to different sampling efficiencies, i.e., asymptotic variances. To utilize multiple statistics for accurate measurements, we formulate a mixture sampling problem, through which we construct a mixture unbiased estimator which minimizes the asymptotic variance. Given fixed sampling budgets for different statistics, we derive the optimal weights to combine the individual estimators; given a fixed total budget, we show that a greedy allocation towards the most efficient statistic is optimal. In practice, the sampling efficiencies of statistics can be quite different for various targets and are unknown before sampling. To solve this problem, we design a two-stage framework which adaptively spends a partial budget to test different statistics and allocates the remaining budget to the inferred best statistic. We show that our two-stage framework is a generalization of 1) randomly choosing a statistic and 2) evenly allocating the total budget among all available statistics, and our adaptive algorithm achieves higher efficiency than these benchmark strategies in theory and experiment.

## I. INTRODUCTION

With the ever increasing popularity of online social networks (OSNs) in recent years, many studies have focused on the analysis of OSNs, such as estimating various properties of the users and their relationships. OSNs are usually measured via graph sampling techniques, because they are typically too large to be completely visited and OSN service providers rarely make their complete network dataset publicly visible. To guarantee the estimation accuracy, many unbiased graph sampling methods have been designed, such as the simple random walk with re-weighting (RWRW) [1, 2], the frontier sampling (FS) [3] and the random walk with uniform restarts (RWuR) [4]. However, OSNs often consist of multiple social graphs which can be sampled by different unbiased graph sampling methods. For example, in the YouTube social network, users are allowed to declare friendship with each other and create interest groups for others to join in. This creates two graphs whose edge sets correspond to 1) the mutual friendship and 2) the sharing of membership of some interest group among the users, respectively. For a given measurement target, sampling via different graphs usually have different efficiencies, which also vary as the measurement target changes. Furthermore, various graph sampling methods can be applied to the same social graph, e.g., the FS and the RWuR are both realiz-

able in the friendship graph of LiveJournal. However, they might induce different sampling efficiencies, which are often unknown a priori. Although one can use multiple unbiased statistics, generated by different methods on different graphs, to form a heterogeneous statistic, it is unclear *how one could* 1) *optimally allocate the sampling budgets among different statistics and* 2) *optimally combine them*.

As we focus on unbiased estimators, we use the asymptotic variance [5] to measure the efficiency of a statistic (or its estimator). We formulate a *mixture sampling problem* that tries to minimize the asymptotic variance of a linearly mixed estimator, constrained by sampling budgets. Given allocated budgets for different statistics, we prove that the optimal weights of individual estimators are inversely proportional to their asymptotic variances; under a fixed total budget, we rank the allocation decisions and find that a greedy allocation is optimal, i.e., allocating more budgets to the statistic with smaller asymptotic variance is always better.

However, the asymptotic variances of the statistics are usually unknown before sampling. To address this challenge, we design a two-stage framework with a pilot and a regular sampling stage. In the pilot sampling stage, we allocate part of the sampling budget to all the statistics and infer the most efficient statistic by estimating the asymptotic variance of each statistic. In the regular sampling stage, we allocate the remaining budget to the inferred most efficient statistic. Our framework is a generalization of two benchmark strategies: 1) spending all budget on a randomly chosen statistic and 2) allocating the budget among all available statistics evenly. We show that our two-stage strategies achieve higher sampling efficiency than the two benchmark strategies. Furthermore, to allocate an optimal sub-budget for the pilot sampling stage, we design an online algorithm to dynamically estimate an upper-bound of the optimal fraction during the pilot sampling. Because the inference of the most efficient statistic is made by estimating the asymptotic variances in the pilot sampling stage, it makes our framework *adaptive* for different measurement targets. Our framework does not restrict how the estimators of asymptotic variances should be constructed, as long as they are asymptotically unbiased. To illustrate, we provide a detailed implementation and evaluate the performance of our framework in the Douban social network. The experimental results show that our technique uses only 18% – 57% of the sampling budget needed by the benchmark strategies

for achieving the same estimation accuracy for a range of measurement targets. Our main contributions are as follows.

- We formulate and solve a mixture sampling problem which constructs an optimal estimator of a heterogeneous statistic to improve sampling efficiency. In particular, we derive the optimal weights of the individual estimators in the mixture estimator (Theorem 1) and the optimal allocation decisions among the statistics (Theorem 2).
- We design a two-stage framework and an adaptive algorithm (Algorithm 1) for the pilot sampling, a practical solution for the mixture sampling problem when the efficiencies of the statistics are unknown before sampling.
- We show that the two-stage strategies are asymptotically optimal (Theorem 4) and achieve higher efficiency than two benchmark strategies (Corollary 1).
- As a case study, we provide a detailed implementation of our framework and evaluate its performance in the Douban social network.

The remaining of this paper is organized as follows. Section II introduces the concepts and characteristics of unbiased graph sampling methods. Section III defines the mixture sampling problem and presents its optimal solution. With unknown efficiencies of the statistics before sampling, we design the two-stage framework and its adaptive algorithm in Section IV. Section V implements the framework and evaluates its performance in the Douban social network. Section VI reviews related work and Section VII concludes.

## II. UNBIASED GRAPH SAMPLING

We denote an undirected graph in an online social network as  $G=(\mathcal{V}, \mathcal{E})$  with a set of nodes  $\mathcal{V}=\{1, \dots, V\}$  to represent users and a set of edges  $\mathcal{E}$  to represent the relationships among the users. We denote  $f$  as a property and  $f_v$  as its value of user  $v$ . Our measurement target is to estimate the mean value of property  $f$  over all users in  $\mathcal{V}$ , i.e.,  $\bar{f} \triangleq (\sum_{v \in \mathcal{V}} f_v) / V$ .

We consider a graph sampling method that traverses the nodes of the graph via a random walk, which generates a discrete-time stochastic process  $\{X_t\}_{t \in \mathbb{N}}$  with the state space of  $\mathcal{V}$ , i.e.,  $X_t \in \mathcal{V}$  for all  $t \in \mathbb{N}$ . We define the random variable  $\hat{f}(m)$  as an estimator on the sample path  $\{X_t: t=1, \dots, m\}$  of  $m$  samples. An estimator  $\hat{f}(\cdot)$  is *unbiased* if  $E[\hat{f}(m)] = \bar{f}$  for all  $m \in \mathbb{N}$  and is *asymptotically unbiased* if

$$\hat{f}(m) \xrightarrow{a.s.} \bar{f} \text{ as } m \rightarrow \infty,$$

where  $\xrightarrow{a.s.}$  denotes *convergence almost surely*. If the process  $\{X_t\}_{t \in \mathbb{N}}$  is ergodic, by the central limit theorem (CLT),

$$\sqrt{m}[\hat{f}(m) - \bar{f}] \xrightarrow{d} N(0, \sigma^2(f)) \text{ as } m \rightarrow \infty, \quad (1)$$

where  $\xrightarrow{d}$  denotes *convergence in distribution* and  $N(0, \sigma^2(f))$  denotes a normal distribution with mean 0 and variance  $\sigma^2(f)$ , which is defined by

$$\sigma^2(f) \triangleq \lim_{m \rightarrow \infty} m \text{Var}(\hat{f}(m)). \quad (2)$$

By (1), we can infer that  $\hat{f}(m) \xrightarrow{a.s.} \bar{f}$  as  $m \rightarrow \infty$ , i.e.,  $\hat{f}(m)$  is an asymptotically unbiased estimator of  $\bar{f}$ . It also shows that the distribution of  $\sqrt{m}\hat{f}(m)$  is asymptotically normal with variance  $\sigma^2(f)$ , which approximately determines how many

samples are required to achieve a certain level of accuracy for the estimator  $\hat{f}(m)$ . Thus, we use the asymptotic variance  $\sigma^2(f)$  to measure the efficiency of an asymptotically unbiased graph sampling method (or its estimator) in this paper.

In the next two sections, we formulate and solve a mixture sampling problem, based on which we design a two-stage framework to sample via multiple statistics. The estimators of these statistics can be based on very different asymptotically unbiased sampling methods on different graphs.

## III. MIXTURE SAMPLING PROBLEM

We consider an objective of measuring the mean value of property  $f$  over the users, i.e.,  $\bar{f}$  defined earlier. We refer to an asymptotically unbiased sampling method on a social graph as a statistic, and assume there are  $K$  types of statistics that can be applied in the OSN. For any statistic  $k$ , we denote the random variable  $\hat{f}_k(m_k)$  as the value of its estimator given  $m_k$  samples and  $\sigma_k^2(f)$  as its asymptotic variance. We simplify the notation  $\sigma_k^2(f)$  as  $\sigma_k^2$  when we focus on a single property  $f$ . Because each estimator  $\hat{f}_k(m_k)$  is asymptotically unbiased, we use the asymptotic variance  $\sigma_k^2$  as a metric for comparing the efficiencies of these statistics. If the asymptotic variance  $\sigma_i^2$  is smaller than  $\sigma_j^2$ , we say statistic  $i$  is more efficient than statistic  $j$  for estimating  $\bar{f}$ . Furthermore, we denote  $k^*$  as the most efficient statistic, i.e.,  $\sigma_{k^*}^2 = \min\{\sigma_k^2 : k = 1, \dots, K\}$ .

### A. Mixture Sampling Problem

Suppose we have a total sampling budget<sup>1</sup> of  $M$  samples and  $K$  types of candidate statistics, we consider the mixture sampling problem of how to allocate the sampling budget among different statistics and how to construct an unbiased estimator  $\hat{f}$  for  $\bar{f}$  so as to minimize its asymptotic variance.

We denote  $\mathbf{a} = (a_1, \dots, a_K)$  as a budget allocation decision, where each  $a_k \geq 0$  defines the fraction of the total budget allocated to statistic  $k$ . We define  $\mathcal{K}_a \triangleq \{k : a_k > 0\}$  to be the set of active statistics. Thus, each active statistic  $k$  has a budget  $m_k = a_k M$  and an estimator  $\hat{f}_k(m_k)$ . Because the sum of budget allocated to each statistic cannot exceed the total budget, we define the constraint set of the allocation decisions as  $\mathcal{A} \triangleq \{\mathbf{a} | \sum_{k=1}^K a_k \leq 1; a_k \geq 0 \forall k = 1, \dots, K\}$ . Given a vector  $\hat{\mathbf{f}} = (\hat{f}_1, \dots, \hat{f}_K)$  of estimators, we consider a mixed estimator  $\hat{f}(\mathbf{w})$  which linearly combines the individual estimators by a weight vector  $\mathbf{w} = (w_1, \dots, w_K)$ , defined as

$$\hat{f}(\mathbf{w}) \triangleq \sum_{k=1}^K w_k \hat{f}_k. \quad (3)$$

Each weight  $w_k$  is used to determine the relative importance of the individual estimator  $\hat{f}_k$ . Under a total budget  $M$  and an allocation decision  $\mathbf{a}$ , we define the *mixture estimator* with weights  $\mathbf{w}$  as

$$\hat{f}(\mathbf{a}, M, \mathbf{w}) \triangleq \sum_{k \in \mathcal{K}_a} w_k \cdot \hat{f}_k(m_k) = \sum_{k \in \mathcal{K}_a} w_k \cdot \hat{f}_k(a_k M). \quad (4)$$

We define the asymptotic variance of the above estimator as

$$\varsigma(\mathbf{a}, \mathbf{w}) \triangleq \lim_{M \rightarrow \infty} M \cdot \text{Var}(\hat{f}(\mathbf{a}, M, \mathbf{w})). \quad (5)$$

<sup>1</sup>We assume that one unit of the budget is the cost of visiting a node.

If each  $\hat{f}_k$  is asymptotically unbiased, we hope that the constructed mixture estimator  $\hat{f}(\mathbf{a}, M, \mathbf{w})$  would still be asymptotically unbiased. We denote the set  $\mathcal{W}_\mathbf{a}$  to be the domain of weights under the budget allocation  $\mathbf{a}$  such that for every  $\mathbf{w} \in \mathcal{W}_\mathbf{a}$ ,  $\hat{f}(\mathbf{a}, M, \mathbf{w})$  is asymptotically unbiased.

Our design goal is to construct the optimal unbiased estimator  $\hat{f}(\mathbf{a}, M, \mathbf{w})$  whose asymptotic variance  $\varsigma(\mathbf{a}, \mathbf{w})$  could be minimized. We formulate two related *mixture sampling problems* as follows. In the first problem, we consider a given allocation decision  $\mathbf{a}$  and we denote  $\varsigma_\mathbf{a}(\mathbf{w}) \triangleq \varsigma(\mathbf{a}, \mathbf{w})$ . The objective is to find the optimal weights  $\mathbf{w}^*$  that solve:

$$\text{Minimize } \varsigma_\mathbf{a}(\mathbf{w}) \quad \text{subject to } \mathbf{w} \in \mathcal{W}_\mathbf{a}. \quad (6)$$

In the second problem, the objective is to find the optimal allocation decision  $\mathbf{a}^*$  and the corresponding optimal weights  $\mathbf{w}^*(\mathbf{a}^*)$  that solve:

$$\begin{aligned} &\text{Minimize} && \varsigma(\mathbf{a}, \mathbf{w}) \\ &\text{subject to} && \mathbf{a} \in \mathcal{A} \quad \text{and} \quad \mathbf{w} \in \mathcal{W}_\mathbf{a}. \end{aligned} \quad (7)$$

The first problem can be regarded as a sub-problem of the second one, where the allocated decision is predetermined.

### B. Optimal Weights and Allocation Decisions

In this subsection, we solve the optimal weights to construct an estimator and the optimal budget allocation for maximizing the efficiency of an estimator. Under a fixed budget allocation decision  $\mathbf{a}$ , intuitively, a larger weight  $w_k$  should be given to an estimator  $\hat{f}_k$  if statistic  $k$  is more efficient, i.e., its asymptotic variance  $\sigma_k^2$  is smaller. The following result provides an affirmative answer to the intuition.

**Theorem 1.** *Assume all the pure estimators  $\hat{f}_k$  are independent of each other. The mixture estimator  $\hat{f}(\mathbf{a}, M, \mathbf{w})$  is asymptotically unbiased for  $\bar{f}$  if and only if the domain of weights under an allocation decision  $\mathbf{a}$  satisfies*

$$\mathcal{W}_\mathbf{a} = \left\{ \mathbf{w} \mid \sum_{k \in \mathcal{K}_\mathbf{a}} w_k = 1 \right\}. \quad (8)$$

*Its asymptotic variance can be characterized by a function of the allocation  $\mathbf{a}$  and the weight vector  $\mathbf{w}$ , defined as*

$$\varsigma(\mathbf{a}, \mathbf{w}) = \sum_{k \in \mathcal{K}_\mathbf{a}} \frac{w_k^2}{a_k} \cdot \sigma_k^2. \quad (9)$$

*The optimal solution  $\mathbf{w}^*$  of the optimization problem in Equation (6) satisfies*

$$w_k^* = \frac{a_k}{\sigma_k^2} / \sum_{i \in \mathcal{K}_\mathbf{a}} \frac{a_i}{\sigma_i^2}, \quad \forall k \in \mathcal{K}_\mathbf{a}, \quad (10)$$

*and the corresponding minimum asymptotic variance is*

$$\varsigma_\mathbf{a}(\mathbf{w}^*) = \left[ \sum_{k \in \mathcal{K}_\mathbf{a}} \frac{a_k}{\sigma_k^2} \right]^{-1}.$$

Theorem 1 shows that to guarantee the mixture estimator to be asymptotically unbiased, the sum of weights of the active statistics must be one. It also tells that when the allocation decision  $\mathbf{a}$  is fixed, the optimal weight  $w_k^*$  of each estimator  $\hat{f}_k(m_k)$  is proportional to  $a_k$  and inversely proportional to its asymptotic variance  $\sigma_k^2$ . Based on Theorem 1, we denote  $\mathbf{w}^*(\mathbf{a})$  to be the optimal solution of (6) defined in (10) and

the second optimization problem (7) could be stated as finding the optimal allocation  $\mathbf{a}^*$  that solves:

$$\text{Minimize } \varsigma(\mathbf{a}, \mathbf{w}^*(\mathbf{a})) \quad \text{subject to } \mathbf{a} \in \mathcal{A}. \quad (11)$$

Intuitively, an optimal solution should allocate more budgets to the more efficient statistic. The next result shows that a greedy strategy that allocates all budgets to the statistic with the smallest asymptotic variance is actually optimal.

**Theorem 2.** *Assume that the conditions of Theorem 1 hold. Denote  $\{\sigma_{(k)}^2\}_{k=1}^K$  as the relabeled set of asymptotic variance of  $\{\sigma_k^2\}_{k=1}^K$  with an ascending order. For any allocation decisions  $\mathbf{a}$  and  $\tilde{\mathbf{a}}$  satisfying  $\sum_{k=1}^i a_{(k)} \geq \sum_{k=1}^i \tilde{a}_{(k)}$  for  $i = 1, 2, \dots, K$ , we have*

$$\varsigma(\mathbf{a}, \mathbf{w}^*(\mathbf{a})) \leq \varsigma(\tilde{\mathbf{a}}, \mathbf{w}^*(\tilde{\mathbf{a}})).$$

*In particular, the optimal allocation  $\mathbf{a}^*$ , which solves the optimization problem in Equation (7), satisfies  $a_k^* = \mathbb{1}_{\{k=k^*\}}$  with the minimum asymptotic variance*

$$\varsigma(\mathbf{a}^*, \mathbf{w}^*(\mathbf{a}^*)) = \sigma_{k^*}^2.$$

Theorem 2 states that an allocation decision  $\mathbf{a}$  is more efficient, i.e., it induces a smaller  $\varsigma(\mathbf{a}, \mathbf{w}^*(\mathbf{a}))$ , if it allocates more budgets to more efficient statistics. In particular, if we greedily allocate all budgets to the most efficient statistic  $k^*$ , the asymptotic variance  $\varsigma(\mathbf{a}, \mathbf{w}^*(\mathbf{a}))$  will be minimized.

Theorem 1 and 2 show that the optimal solutions are closely related to the asymptotic variances  $\sigma_k^2$  of the individual statistics, and the directions for decreasing  $\varsigma(\mathbf{a}, \mathbf{w})$  are allocating as much budget to statistic  $k^*$  as possible and weighting the individual estimators inversely proportional to their asymptotic variances. However, the asymptotic variances  $\sigma_k^2$  are usually unknown before sampling. To address this challenge, we propose a two-stage framework, where we infer the best statistic  $k^*$  in the first stage before allocating all the remaining budget greedily in the second stage.

### IV. ADAPTIVE TWO-STAGE FRAMEWORK

In this section, we first explain the basic concepts of a two-stage framework and then show the framework achieves higher sampling efficiency than two benchmark strategies, finally we propose an adaptive algorithm to determine an upper-bound of the optimal budget fraction which is allocated to the first stage.

#### A. Two Benchmark Strategies and A Two-Stage Generalization

Without knowing the asymptotic variances  $\sigma_k^2$  of the individual statistics, we start with two naive strategies as benchmarks. The first strategy spends all budget  $M$  on a randomly chosen statistic  $k$ ; the second strategy evenly divides the budget  $M$  among  $K$  statistics to construct the mixture estimator. We call these two benchmark strategies as the *Random Statistics* (or RND) and *Average Statistics* (or AVG), respectively.

Based on the two benchmark strategies, we consider a two-stage generalization, which spends a partial budget to estimate the best statistic  $k^*$  in a *pilot sampling stage* and allocates the remaining budget to an estimated best statistic  $\hat{k}^*$  in a *regular sampling stage*. We assume that a fraction  $c \in [0, 1]$  of the

total budget  $M$  is allocated for pilot sampling and name the  $cM$  samples as the pilot budget. We evenly allocate the pilot budget among all  $K$  statistics, and therefore, each statistic  $k$  is allocated a budget of  $m_k = cM/K$  samples in this stage. We use these pilot samples to make an *asymptotically unbiased estimate* of each asymptotic variance  $\sigma_k^2$ , and define the estimated value by  $\hat{\sigma}_k^2(m_k)$ . Most likely, the statistic with the smallest estimated asymptotic variance tends to be the most efficient statistic  $k^*$  for estimating  $\bar{f}$ . We call this statistic the *inferred most efficient statistic* and denote it as  $\hat{k}^*(cM)$ , parameterized by the pilot sampling budget  $cM$ . In the regular sampling stage, we allocate all the remaining sampling budget  $(1-c)M$  to the inferred most efficient statistic  $\hat{k}^*$ , and fully use the total budget  $M$  to construct a mixture estimator.

Under the above two-stage framework, we denote  $\mathbf{a}(c)$  as the effective allocation decision, defined by

$$a_k(c) \triangleq c/K + (1-c) \cdot \mathbb{1}_{\{k=\hat{k}^*(cM)\}}, \quad (12)$$

through which we can define the effective budget for each statistic  $k$  as  $m_k(c) \triangleq a_k(c)M$  naturally. After both sampling stages, we construct a mixture estimator by using an estimated optimal weight vector  $\hat{\mathbf{w}}^*(c)$ . We use the estimated value  $\hat{\sigma}_k^2(m_k)$  to approximate  $\sigma_k^2$ , and define  $\hat{\mathbf{w}}^*(c)$  by substituting  $\sigma_k^2$  with  $\hat{\sigma}_k^2(m_k)$  in the optimal weight of Equation (10) as

$$\hat{\mathbf{w}}^*(c) \triangleq \frac{a_k(c)}{\hat{\sigma}_k^2(m_k(c))} / \sum_{i \in \mathcal{K}_a} \frac{a_i(c)}{\hat{\sigma}_i^2(m_i(c))}, \quad \forall k \in \mathcal{K}_a. \quad (13)$$

Consequently, the corresponding mixture estimator and its asymptotic variance can be written as  $\hat{f}(\mathbf{a}(c), M, \hat{\mathbf{w}}^*(c))$  and  $\varsigma(\mathbf{a}(c), \hat{\mathbf{w}}^*(c))$ , respectively.

The two-stage framework actually uses the AVG and RND strategies in its pilot and regular sampling stages, respectively. In particular, the estimated statistic  $\hat{k}^*$  plays the role of a random statistic in the RND strategy. Also, the framework can be seen as a generalization of the two benchmark strategies, because the Average and Random Statistics are equivalent to a two-stage strategy of  $c = 1$  and  $c = 0$ , respectively.

**Theorem 3.** *The asymptotic variances of the Random Statistics and Average Statistics are  $\varsigma(\mathbf{a}(0), \hat{\mathbf{w}}^*(0))$  and  $\varsigma(\mathbf{a}(1), \mathbf{a}(1))$ , respectively. They satisfy*

$$\mathbb{E}[\varsigma(\mathbf{a}(0), \hat{\mathbf{w}}^*(0))] = \varsigma(\mathbf{a}(1), \mathbf{a}(1)) = \frac{1}{K} \sum_{k=1}^K \sigma_k^2.$$

Theorem 3 states that the expected asymptotic variance of the Random Statistics and the asymptotic variance of Average Statistics both equal the average of the asymptotic variances of all individual statistics.

### B. Asymptotic Performance of Two-Stage Strategies

Our two-stage framework does not restrict how the asymptotic variances  $\sigma_k^2$  are estimated in the pilot sampling stage. We will show that as long as  $\hat{\sigma}_k^2(\cdot)$  is an asymptotically unbiased estimator for  $\sigma_k^2$ , the two-stage strategies will outperform the two benchmark strategies. The detailed design of the estimator  $\hat{\sigma}_k^2(\cdot)$  may depend on the sampling method of statistic  $k$ , and we will give an example of implementation in a later section.

Given any strategy  $c \in [0, 1]$ , we can define the (unknown) optimal allocation decision as  $\mathbf{a}^*(c) = (a_1^*(c), \dots, a_K^*(c))$  as

$$a_k^*(c) \triangleq c/K + (1-c) \cdot \mathbb{1}_{\{k=k^*\}}, \quad \forall k = 1, \dots, K.$$

Under this optimal allocation  $\mathbf{a}^*(c)$ , by Theorem 1, the corresponding optimal weight vector becomes  $\mathbf{w}^*(\mathbf{a}^*(c))$ . Intuitively, when a budget  $cM$  is used to estimate each  $\sigma_k^2$  in the pilot sampling stage,  $m_k = cM/K$  for any statistic  $k$  and the best statistic  $k^*$  is more likely to induce a smaller estimated asymptotic variance  $\hat{\sigma}_k^2(m_k)$  than other statistics. Consequently, the resulting allocation  $\mathbf{a}(c)$  and weights  $\hat{\mathbf{w}}^*(c)$  are more likely to be equal to the optimal  $\mathbf{a}^*(c)$  and  $\mathbf{w}^*(\mathbf{a}^*(c))$ , respectively. We consider the two-stage strategy  $c$  as a function of the total budget  $M$ , denoted as  $c(M)$ , and simplify the notation  $\mathbf{a}^*(c(M))$  as  $\mathbf{a}^*(M)$ . The next theorem shows that when the pilot budget fraction  $c$  is higher than the order of  $M^{-1}$ , the two-stage strategy  $c(M)$  is asymptotically optimal.

**Theorem 4.** *Assume each estimated asymptotic variance  $\hat{\sigma}_k^2(\cdot)$  is asymptotically unbiased for  $\sigma_k^2$  ( $k = 1, \dots, K$ ), i.e.,*

$$\hat{\sigma}_k^2(m_k) \xrightarrow{a.s.} \sigma_k^2 \quad \text{as } m_k \rightarrow +\infty.$$

*If  $c(M) \in \omega(M^{-1})$ , i.e., for all  $\delta > 0$ , there exists a positive number  $M'$  such that  $c(M) \geq \delta M^{-1}$  for all  $M > M'$ ,*

$$\hat{k}^* \xrightarrow{a.s.} k^*, \quad \mathbf{a}(c(M)) \xrightarrow{a.s.} \mathbf{a}^*(M) \quad \text{and}$$

$$\hat{\mathbf{w}}^*(c(M)) \xrightarrow{a.s.} \mathbf{w}^*(\mathbf{a}^*(M)) \quad \text{as } M \rightarrow +\infty.$$

Theorem 4 shows that as the total budget  $M$  grows, to guarantee an (asymptotic) optimal two-stage strategy, the fraction  $c$  for the pilot budget does not need to be large. The condition  $c(M) \in \omega(M^{-1})$  ensures that the pilot budget  $c(M)M$  grows with  $M$  unboundedly as  $M$  goes to infinity, although  $c$  itself could approach zero, such that the estimated asymptotic variance  $\hat{\sigma}_k^2(m_k)$  will converge to  $\sigma_k^2$ . Consequently, the two-stage strategy  $c(M)$  will identify the most efficient statistic  $k^*$  via the pilot sampling and set the optimal allocation  $\mathbf{a}^*(c(M))$  and optimal weight  $\mathbf{w}^*(\mathbf{a}^*(c(M)))$  for the mixture estimator.

When we simply give the same weight for each sample point, for any allocation  $\mathbf{a}$ , the corresponding weight vector becomes  $\mathbf{w} = \mathbf{a}$ , which are proportional to their sample sizes. To distinguish the benefit of choosing an optimal allocation  $\mathbf{a}^*$  and an optimal weight  $\mathbf{w}^*$ , we consider an intermediate mixture estimator  $\hat{f}(\mathbf{a}, M, \mathbf{a})$ , which gets affected only by the allocation decision  $\mathbf{a}$  and has an asymptotic variance  $\varsigma(\mathbf{a}, \mathbf{a})$ .

**Corollary 1.** *Under the conditions of Theorem 4, for any pilot fraction  $c(M) \in \omega(M^{-1})$ , as  $M \rightarrow +\infty$ , we have*

$$\varsigma(\mathbf{a}(c(M)), \hat{\mathbf{w}}^*(c(M))) \xrightarrow{a.s.} \varsigma(\mathbf{a}^*(M), \mathbf{w}^*(\mathbf{a}^*(M))),$$

*and the asymptotic limit of  $\varsigma$  satisfies*

$$\varsigma(\mathbf{a}^*(M), \mathbf{w}^*(\mathbf{a}^*(M))) \leq \varsigma(\mathbf{a}^*(M), \mathbf{a}^*(M)) \leq \frac{1}{K} \sum_{k=1}^K \sigma_k^2$$

$$\text{and } \varsigma(\mathbf{a}^*(M), \mathbf{w}^*(\mathbf{a}^*(M))) \leq \frac{K\sigma_{k^*}^2}{K + (1-K)c(M)}.$$

As a consequence of Theorem 4, Corollary 1 shows that as  $M$  grows, the asymptotic variance  $\varsigma$  induced by the strategy  $c(M)$  converges to an optimal value  $\varsigma(\mathbf{a}^*(M), \mathbf{w}^*(\mathbf{a}^*(M)))$ .

The first inequality implies that 1) using the estimated optimal weight  $\hat{w}^*(c(M))$  is more efficient than the equal weight  $w = \mathbf{a}$ , and 2) using  $w = \mathbf{a}$  is again more efficient than the two benchmark strategies, whose (expected) asymptotic variances equal  $\frac{1}{K} \sum_{k=1}^K \sigma_k^2$  as shown in Theorem 3. The second inequality provides an upper-bound for the optimal  $\varsigma$ , which can be derived from an estimator  $\hat{f}_{\hat{k}^*}(a_{\hat{k}^*}(c(M))M)$  which only uses the samples of the inferred best statistic  $\hat{k}^*$  and throws out the samples of other statistics collected in the pilot sampling stage.

### C. Optimal Fraction for Pilot Budget

Our design of any two-stage strategy  $c(M) \in \omega(M^{-1})$  is asymptotically optimal. However, a more practical problem is that, given a finite budget  $M$ , how to choose an optimal fraction  $c^*(M)$  for the pilot budget that maximizes the efficiency for the mixture estimator  $\hat{f}$ , i.e.,  $c^*(M)$  solves:

$$\begin{aligned} & \text{Minimize} && \text{Var}(\hat{f}(\mathbf{a}(c), M, \hat{w}^*(c))), \\ & \text{subject to} && c \in [0, 1]. \end{aligned} \quad (14)$$

On the one hand, when allocating more budget for the pilot sampling, each  $\hat{\sigma}_k^2(\cdot)$  could provide a more accurate estimation for the asymptotic variance  $\sigma_k^2$  and the best statistic  $k^*$  would have a higher chance to be picked out in the regular sampling stage. On the other hand, increasing the pilot budget means that more budget will be allocated to some inefficient statistics at the pilot sampling stage. One needs to balance the above contradictory conditions so as to obtain an optimal fraction  $c^*(M)$ . In practice, it is hard to obtain the exact value of the optimal fraction for the pilot budget  $c^*(M)$ , because it depends on the unknown values of asymptotic variances  $\sigma_k^2$ . However, we will provide a heuristic algorithm to estimate  $c^*(M)$  effectively, which is based on the following theoretical result on the monotonicity of  $c^*(M)$ .

**Theorem 5.** Assume the rate of convergence of the estimated asymptotic variance  $\hat{\sigma}_k^2(m)$  for each  $\sigma_k^2$  is  $\Theta(m^{-\eta_k})$ , i.e.,  $\sup_{x \in \mathbb{R}^+} |G_{\hat{\sigma}_k^2(m)}(x) - G_{\sigma_k^2}(x)| = \Theta(m^{-\eta_k})$ , where  $G_{\hat{\sigma}_k^2(m)}(x) = \mathbb{P}(\hat{\sigma}_k^2(m) \leq x)$  and  $G_{\sigma_k^2}(x) = \mathbb{1}_{\{x \geq \sigma_k^2\}}$  are the cumulative distribution function of  $\hat{\sigma}_k^2(m)$  and  $\sigma_k^2$ , respectively, and the order  $\eta_k > 0$ . Let  $\eta = \min\{\eta_k : k = 1, \dots, K\}$ . The optimal fraction satisfies  $\lim_{M \rightarrow +\infty} c^*(M) = 0$  with the rate of convergence  $\Theta(M^{-1+\frac{1}{\eta+1}})$ .

Theorem 5 shows that the optimal fraction  $c^*(M)$  decreases to zero asymptotically with the rate  $\Theta(M^{-1+\frac{1}{\eta+1}})$  when  $M$  grows. Intuitively, as the total budget  $M$  increases, to guarantee the same accuracy for estimating  $k^*$ , we only need to keep the pilot budget  $cM$  constant and thus the fraction  $c$  becomes smaller. Both Theorem 4 and 5 imply that when  $M$  becomes larger, the optimal fraction  $c^*(M)$  should decrease. Therefore, we assume that  $c^*(M)$  follows a decreasing trend as  $M$  increases (In Section V, our evaluations in the Douban social network also support this conjecture well), based on which we propose an adaptive algorithm to dynamically determine the optimal fraction  $c^*(M)$  for the pilot sampling.

---

### Algorithm 1 Adaptive Two-Stage Sampling ( $M, \Delta_M$ )

---

- 1:  $c \leftarrow \Delta_M/M$ ;
  - 2: spend  $\Delta_M$  budget for pilot sampling;
  - 3: **while**  $c < \hat{c}^*(cM)$  **do**  
 $c \leftarrow c + \Delta_M/M$ ;  
 spend  $\Delta_M$  more budget for pilot sampling;  
**end while**
  - 4: choose the estimated best statistic  $\hat{k}^*$ ;
  - 5: spend the remaining budget  $(1-c)M$  for regular sampling;
- 

Algorithm 1 performs the pilot sampling in an adaptive manner. It takes two input parameters: the total budget  $M$  and a budget spending stepsize  $\Delta_M \in (0, M)$ . We denote  $\hat{c}^*(\cdot)$  as a function where each  $\hat{c}^*(m)$  provides an estimated upper-bound of the optimal fraction  $c^*(M)$ , when  $m$  number of samples are used. In step 3, we increase the pilot budget by  $\Delta_M$  if the spent fraction  $c$  is smaller than the derived upper-bound  $\hat{c}^*(cM)$  for  $c^*(M)$ , until  $c$  exceeds the upper-bound  $\hat{c}^*(cM)$ . Based on the  $cM$  samples generated in the pilot sampling stage, we choose the estimated best statistic  $\hat{k}^*$  and spend the remaining budget  $(1-c)M$  for regular sampling as usual. In general, given any  $m$  pilot samples, the function  $\hat{c}^*(\cdot)$  uses them to estimate an optimal fraction  $c^*(m')$  for some  $m' < m$ . Because  $c^*(\cdot)$  has a decreasing trend in general, we could use this estimation of  $c^*(m')$  as an upper-bound for  $c^*(M)$  so as to determine whether the pilot sampling stage should end. As the sampling budget  $cM$  increases, the estimation  $\hat{c}^*(cM)$  should decrease and approach  $c^*(M)$ , because it estimates some  $c^*(m')$  and  $m'$  increases. Notice that our algorithm does not restrict how the upper-bound estimation  $\hat{c}^*(\cdot)$  should be implemented, and we will provide an example of implementation which we use in our evaluation in a later section. Finally, although a large stepsize approaches  $c^*(M)$  faster, to avoid overestimating the pilot budget, a small value of  $\Delta_M$  should be used in practice.

## V. EVALUATION IN DOUBAN SOCIAL NETWORK

In this section, we apply the adaptive two-stage framework to the Douban social network, a popular Chinese web site providing user comment and recommendation services for books, music and movies. We first introduce multiple statistics which can be realized in Douban and then provide a detailed implementation of our framework to measure the statistics, finally we evaluate the performance of the framework.

### A. Multiple Statistics

Similar to Twitter and Sina microblog, users in Douban can follow each other, and therefore, Douban can be seen as a followship graph<sup>2</sup> in which the edges capture the following relationship. Douban also allows users create interest groups for others to join in. We consider two users who have a common group share a membership and Douban can also be seen as a membership graph. These two different social graphs,

<sup>2</sup>Here, we serve the followship graph as an undirected graph and one following relationship corresponds to an undirected edge.

together with two random walk based sampling methods, the RWuR and FS introduced next, provide four different available statistics.

1) *The random walk with uniform restarts (RWuR)*: The RWuR [4] is a hybrid sampling method that mixes random walk crawling and uniform node sampling. It generates a sample set  $\{X_t\}_{t \in \mathbb{N}}$  as follows. At each step  $t$ , assume the current node is  $X_t = i$ . With probability  $\alpha/(d_i + \alpha)$ , it jumps to an arbitrary node  $j$  of the graph chosen uniformly and make the transition  $X_{t+1} = j$ . With probability  $d_i/(d_i + \alpha)$ , it uniformly chooses an  $i$ 's neighboring node  $k$ , i.e.,  $X_{t+1} = k$ . The parameter  $\alpha$  ( $\geq 0$ ) controls the probabilities of random walk and jump. Specially, when  $\alpha = 0$ , the RWuR is the simple random walk, and when  $\alpha = +\infty$ , the RWuR becomes the uniform node sampling. Obviously, the sample set  $\{X_t\}_{t \in \mathbb{N}}$  is biased towards the high-degree nodes. To correct the bias, it uses the Hansen-Hurwitz estimator [6, 7] to re-weight the samples, i.e., the weight of the sample  $X_t$  is inversely proportional to  $d_{X_t} + \alpha$ , and the unbiased estimator for  $\bar{f}$  is

$$\hat{f}(m) = \sum_{t=1}^m \frac{f_{X_t}}{d_{X_t} + \alpha} / \sum_{t=1}^m \frac{1}{d_{X_t} + \alpha}. \quad (15)$$

2) *The frontier sampling (FS)*: The FS [3] is a distributed sampling method that performs  $s$  ( $\in \mathbb{N}$ ) random walkers on a graph. Initially, it uniformly obtains  $s$  nodes as the start nodes of the  $s$  random walkers. At each step, it first randomly selects the  $r$ -th walker with probability  $d_{v_r} / \sum_{i=1}^s d_{v_i}$ , where  $v_i$  is the current node of the  $i$ -th walker. Then the  $r$ -th walker uniformly chooses a  $v_r$ 's neighboring node as the next sample and moves to it. Similar to the RWuR, the bias towards high degree nodes of the FS can be corrected by the Hansen-Hurwitz estimator, and the unbiased estimator for  $\bar{f}$  is

$$\hat{f}(m) = \sum_{t=1}^m \frac{f_{X_t}}{d_{X_t}} / \sum_{t=1}^m \frac{1}{d_{X_t}}. \quad (16)$$

The RWuR and FS samplers are less likely to get trapped in loosely connected components of a graph via jumping randomly and running multiple walkers, respectively. Thus, both of them usually perform better than the simple random walk with re-weighting [1, 2], but we do not know which one achieves higher sampling efficiency in an unknown graph. Besides, it is unclear how the efficiencies of the two methods vary on the followship and membership graphs. Therefore, we choose the four statistics, the RWuR and FS on the followship and membership graphs, to demonstrate our two-stage framework.

## B. Implementation of Two-Stage Framework

Our adaptive two-stage strategy does not restrict how the estimators of the asymptotic variances  $\hat{\sigma}_k^2(\cdot)$  and the upper-bound estimation of the optimal fraction  $\hat{c}^*(\cdot)$  are constructed, as long as they are asymptotically unbiased. Next, we provide an example of detailed implementations of  $\hat{\sigma}_k^2(\cdot)$  and  $\hat{c}^*(\cdot)$  for measuring in Douban.

1) *Estimating the asymptotic variances*: Both the pilot sampling and the adaptive Algorithm 1 need to estimate the unknown asymptotic variances  $\sigma_k^2$ . Assume the sample set used to estimate  $\sigma_k^2$  is collected by  $q$  ( $\geq 2$ ) samplers whose budgets are all  $l$ . We denote the estimated value for  $\bar{f}$  based on the  $j$ -th sampler as  $\hat{f}_k^{(j)}(l)$  for  $j = 1, \dots, q$ , which serves as a sample of the estimator  $\hat{f}_k(l)$ . Then the sample variance of  $\{\sqrt{l}\hat{f}_k^{(j)}(l) : j = 1, \dots, q\}$  is defined by

$$S_k^2(q, l) = \frac{l}{q-1} \sum_{j=1}^q \left[ \hat{f}_k^{(j)}(l) - \frac{1}{q} \sum_{i=1}^q \hat{f}_k^{(i)}(l) \right]^2. \quad (17)$$

It describes how far  $\sqrt{l}\hat{f}_k^{(j)}(l)$  ( $j = 1, 2, \dots, q$ ) are spread out. From the definition of asymptotic variance in Equation (2),  $S_k^2(q, l)$  is an asymptotically unbiased estimate of  $\sigma_k^2$ , i.e.,

$$S_k^2(q, l) \xrightarrow[l \rightarrow \infty]{a.s.} \lim_{l \rightarrow \infty} l \text{Var}(\hat{f}_k(l)) = \sigma_k^2 \quad \text{as } q, l \rightarrow \infty. \quad (18)$$

Thus, we can use  $S_k^2(q, l)$  to estimate the asymptotic variance  $\sigma_k^2$ , i.e.,  $\hat{\sigma}_k^2(q, l) = S_k^2(q, l)$ . Also, because all unbiased graph sampling methods have the same definition of the asymptotic variance from Equation (2), this implementation is applicable to any one of them.

2) *Estimating upper-bound of optimal fraction  $c^*(M)$* : Given any sub-budget  $m < M$ , we provide an implementation of the upper-bound estimation function  $\hat{c}^*(m)$  as follows. We use the budget  $m$  to collect  $B$  sample sets whose sizes are all  $m' \triangleq \frac{m}{BK}$  for each statistic, and by using these  $m$  samples, we could estimate the optimal fraction  $c^*(m')$  when the total given budget is  $m'$ . We denote the  $b$ -th sample set of the  $k$ -th statistic as  $S_k^{(b)}$ . We could try different two-stage strategies with fraction  $c \in (0, 1)$  on the  $b$ -th group of sample sets  $\{S_k^{(b)} : k = 1, \dots, K\}$ . Specifically, like a normal two-stage strategy of fixed  $c$ , we obtain  $\frac{cm'}{K}$  samples from each set  $S_k^{(b)}$  as the pilot sampling, and use them to estimate the best statistic  $k^*$ , and then use the remaining  $(1-c)m'$  samples of the inferred best statistic to generate a realization of the estimator  $\hat{f}(a(c), m', \hat{w}^*(c))$ . Finally, we calculate the sample variance of the  $B$  realizations obtained from the  $B$  groups of sample sets. Based on (14), we choose the fraction  $c$  that minimizes the sample variance as an estimation for  $c^*(m')$ , which serves as an upper-bound for the optimal fraction  $c^*(M)$ .

When increasing the number of realizations  $B$ , the estimation  $\hat{c}^*(m')$  for  $c^*(m')$  becomes more accurate. However, the budget  $m' = \frac{m}{BK}$  decreases under a fixed budget  $m$ , and therefore, using  $\hat{c}^*(m')$  as an upper-bound for the optimal fraction  $c^*(M)$  could be loose. As a result, we recommend to set the parameter  $B$  moderately.

## C. Measurement Setup

The publicly available information for every Douban user includes user-id, location, lists of followers, users he/she follows and the interest groups he/she joins in. We consider two measurement targets, i.e., the average number of followers of users and the average number of interest groups of users. To measure these targets, we develop crawlers to sample via the four statistics ( $K = 4$ ), i.e., the RWuR and FS methods on the followship and membership graphs. We ignore the users who

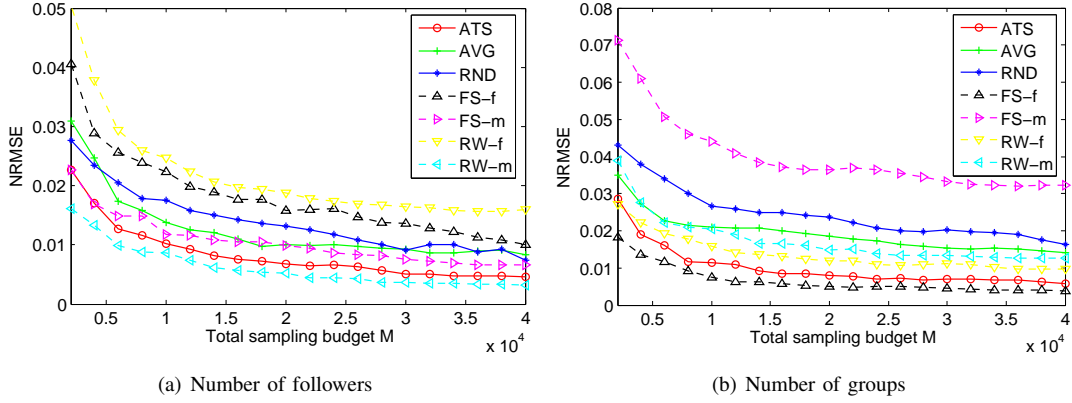


Fig. 1. NRMSE of the adaptive two-stage strategy (ATS), Average Statistics (AVG), Random Statistics (RND), and individual statistics including the RWuR and FS methods on the followership graph (RW-f and FS-f) and on the membership graph (RW-m and FS-m), when we vary the total sampling budget  $M$ . (a) measure the average number of followers of users, and (b) measure the average number of interest groups of users.

do not have any followership or membership, as these isolated users cannot be visited via crawling.

We set the total sampling budget to be  $M = 4 \cdot 10^4$ , which represents about 0.05% of the total number of Douban users<sup>3</sup>. For the statistics based on the FS method, we set the number of random walkers  $s = 50$  in a FS sampler. For the statistics based on the RWuR method, we moderately set the parameter which controls the probabilities of random walk and jump  $\alpha = 0.1$ . Besides, we also consider the cost of uniformly choosing a start node and jumping to an arbitrary node in the FS or RWuR method. This cost is about 14 units of budget in the Douban network, i.e., it needs to query an average of 14 randomly generated user-ids to obtain a valid one in the user-id space. In the two-stage framework, we set the number of realizations  $B = 10$  and the budget spending stepsize  $\Delta_M = 2\%M = 800$  for Algorithm 1. To estimate the asymptotic variances, we use  $q = 5$  samplers.

We also implement the benchmark strategies, i.e., the Random Statistics and Average Statistics, and the four single-statistic strategies for comparison. To measure the estimation accuracy of the different sampling strategies, we use Normalized Root Mean Square Error (NRMSE) [2–4],  $\sqrt{\mathbb{E}(\hat{f} - \bar{f})^2 / \bar{f}}$  where  $\bar{f}$  is the true value of the measurement target and  $\hat{f}$  is the estimated one. Because the “ground truth”  $\bar{f}$  is not published by Douban, we calculate the NRMSE by taking as  $\bar{f}$  the grand average of  $\hat{f}$  values over all samples collected via all full-length crawlers and statistics. All experiment results presented in the following are the average of 25 independent simulations and our crawls were performed from Nov. 5th to 11th of 2013.

#### D. Evaluation Results

##### 1) Performance of the Adaptive Two-Stage Strategies (ATS):

Figure 1 shows that the efficiencies of different statistics may vary for different measurement targets. For example, we observe that when we measure the average number of followers

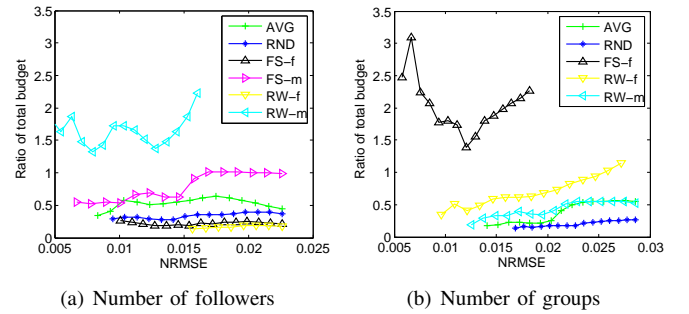


Fig. 2. The ratio of needed budget between the two-stage strategy and others including the Average Statistics (AVG), Random Statistics (RND), and individual statistics including the RWuR and FS methods on the followership graph (RW-f and FS-f) and on the membership graph (RW-m and FS-m), so as to attain the same NRMSE. (a) measure the average number of followers of users, and (b) measure the average number of interest groups of users.

of users, the RWuR method on the membership graph (RW-m) leads to higher estimation accuracy than the FS method on the followership graph (FS-f) as shown in subfigure 1(a); however, when the target is the average number of interest groups of users, the conclusion is reversed as shown in subfigure 1(b). Thus, the efficiencies of the statistics vary as the measurement target changes and choosing a bad statistic, e.g., the RW-f strategy for estimating the average number of followers, may lead to an inaccurate estimation. Without knowing the efficiencies of the individual statistics, Figure 1 shows that our adaptive two-stage strategy (ATS) always outperforms both benchmark strategies (AVG and RND) regardless of the measurement target. Furthermore, our strategy (ATS) is only a bit inferior to the true best statistic (RW-m for estimating the average number of user’s followers or FS-f for estimating the average number of user’s groups), which could be used when the asymptotic variances of all statistics are known. Figure 1 also demonstrates that our framework has good adaptivity for different measurement targets in the two subfigures.

Figure 2 shows the budget saving of our two-stage strategy (ATS) compared with the benchmarks (AVG and RND) and other single-statistic strategies if they can fulfill the given NRMSE target. For example, when measuring the average

<sup>3</sup>By Nov. 15 2013, Douban service provider declare there are about 79.2 million users.



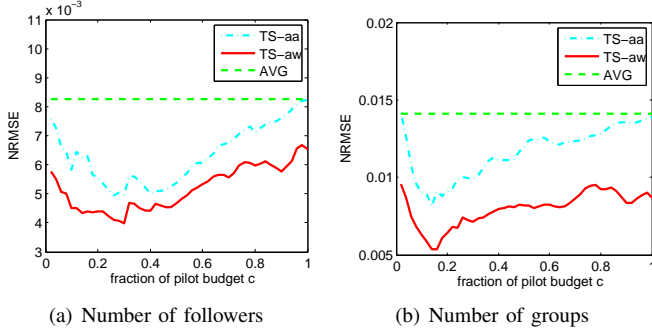


Fig. 3. With total sampling budget  $M = 4 \cdot 10^4$ , NRMSE of the two-stage framework with the estimated optimal weights, i.e.,  $\hat{f}(\mathbf{a}, M, \hat{\mathbf{w}}^*)$  (TS-aw), the two-stage framework with the same weight for each sample point, i.e.,  $\hat{f}(\mathbf{a}, M, \mathbf{a})$  (TS-aa) and Average Statistics (AVG), when we vary the fraction of pilot budget  $c$ . (a) measure the average number of followers of users, and (b) measure the average number of interest groups of users.

number of followers of users in subfigure 2(a), ATS saves about 49% budget compared with the AVG strategy to obtain the  $\text{NRMSE} = 0.015$ . From subfigures 2(b), as the target is the average number of groups of users, ATS saves about 75% budget compared to the RND strategy for obtaining the  $\text{NRMSE} = 0.025$ . In general, we observe that our ATS strategy requires only 18% to 57% of the budget needed for the benchmark strategies to achieve the same NRMSE. We also observe when measuring the average number of followers (resp. groups) of users, the best statistic RW-m (resp. FS-f) uses the smallest amount of budget, which is consistent with the observations from Figure 1 and the result of Theorem 2.

2) *Benefit of optimal allocation decision and weights*: The two-stage framework tries to improve estimation efficiency by choosing budget allocation decision and setting estimated optimal weights for the mixture estimator. Figure 3 compares the NRMSE of our two-stage strategy when the weights are set to be equal (TS-aa) or optimally adjusted (TS-aw) and that of the AVG benchmark strategy, when the fraction  $c$  of the pilot budget varies along the x-axis. We observe that the two-stage strategy with optimal weights always outperforms that with equal weights, which again outperforms the AVG benchmark strategy. Notice that under the equal weights,  $c = 0$  and  $c = 1$  corresponds to the RND and AVG strategies, respectively, which have the same performance as shown in Theorem 3. In general, when  $c$  increases from 0 to 1, the benefit of two-stage strategy first increases and then decreases. This is an integrated result of two competing factors: 1) increasing the pilot budget help select the more efficient statistic at the regular sampling stage, and 2) at the same time more budgets are allocated to the inefficient statistics at the pilot sampling stage. We also observe that the benefit of using optimal weights is larger when the pilot fraction is larger. The reason is that with the larger pilot budget, more samples are used on the inefficient statistics and therefore, optimal weights are more needed to discount those statistics.

3) *Effectiveness of the adaptive Algorithm 1*: We implemented Algorithm 1 for estimating the optimal pilot fraction. Figure 4 shows that the estimated optimal pilot fraction

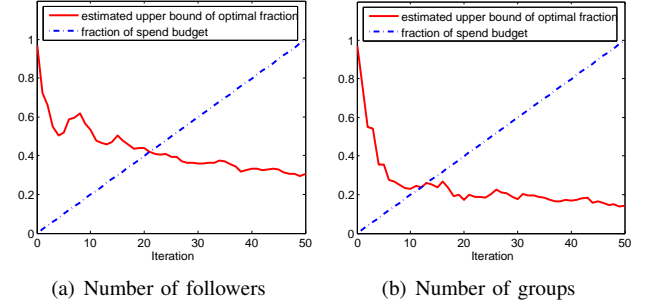


Fig. 4. With total sampling budget  $M = 4 \cdot 10^4$ , the estimated upper bound of the optimal fraction  $\hat{c}^*(cM)$  and the spent fraction of budget  $c$  when the iteration increases in Algorithm 1. (a) measure the average number of followers of users, and (b) measure the average number of interest groups of users.

$\hat{c}^*(t\Delta_M)$  for  $c^*(M)$  (solid line) has a decreasing trend as the number of iterations  $t$  increases. It is consistent with our result that the optimal pilot fraction  $c^*(M)$  decreases as the budget  $M$  grows. The consumed fraction of the pilot budget  $c$  (dash line) increases linearly (at a rate of  $\Delta_M$ ) with the number of iterations. When the consumed pilot fraction  $c$  is larger than the estimated upper-bound of optimal fraction, the iteration stops in Algorithm 1. Subfigures 4(a) (resp. 4(b)) show that when measuring the average number of users' followers (resp. groups), the estimated optimal pilot fraction 40% (resp. 22%) approximates efficiently the real value 32% (resp. 16%). These results show that Algorithm 1 is effective for setting a near-optimal pilot fraction in the practical two-stage sampling.

4) *Observations of different statistics*: At last, we provide some insights into the different statistics. Subfigure 1(a) indicates that the RWuR and FS methods on the membership graph perform better than them on the followship graph when we measure the average number of followers of users. However, when the target is the average number of groups of users, the conclusion is reversed as shown in subfigure 1(b). The reason may be that the followship (resp. membership) graph has a strong cluster feature [8] that makes the samples highly correlated on the number of the users' followers (resp. groups). This strong correlation leads to a poor estimation accuracy. We also observe that, for the followship graph, the FS method achieves higher efficiency than the RWuR; while the RWuR has smaller estimation error than the FS for the membership graph. Because the RWuR sampler frequently chooses an arbitrary node as restart on a less connected graph (e.g., the followship graph), which costs large budget and decreases the estimation accuracy. On the other hand, the RWuR is close to a single random walker on a well connected graph (e.g., the membership graph). Compared with the FS with multiple random walkers, it saves the cost of obtaining multiple uniform start nodes and converging to the walkers' steady state.

## VI. RELATED WORK

*Graph Sampling Techniques*. As OSN service providers rarely make publicly visible to the frame information of entire networks, most widely used graph sampling techniques in OSNs are crawling methods. Early graph crawling methods



are based on Breath-First Search (BFS), Depth-First Search (DFS) and Snowball Sampling (SBS) [9]. In particular, BFS has been frequently used to explore large networks, such as Youtube and Facebook [6]. However, these methods introduce a large bias towards high degree nodes and it is difficult to be corrected in general graphs [10–13].

Recently the most popular graph crawling is random walk-based sampling, including simple random walk with re-weighting (RWRW) [1, 2] and Metropolis-Hastings random walk (MHRW) [14]. RWRW is considered as a special case of Respondent-Driven Sampling (RDS) [1] if only one neighbor is chosen in each iteration and revisiting nodes is allowed. It is also biased to sample high degree nodes, but the bias can be corrected by the Hansen-Hurwitz estimator shown in [6, 7]. RWRW was not only used to sample OSNs [7, 12], but also P2P networks and Web [15, 16]. MHRW is based on the Metropolis-Hastings (MH) algorithm and provides unbiased samples directly [2, 14]. Some studies [1, 2] have shown that RWRW estimates are more accurate than MHRW estimates.

*Improvement of sampling efficiency.* Researchers have proposed some methods to improve the sampling efficiency against random walk-based sampling, including the FS [3] and RWuR [4] methods which we apply as showcases in this work. Besides, Kurant et al. [17] presented a weighted random walk method to perform stratified sampling with a priori estimate of network information. Lee et al. [18] proposed a non-backtracking random walk which forbids the sampler to backtrack to the previously visited node, and they theoretically guaranteed the technique achieves higher efficiency than a simple random walk. Our work concentrates on how to combine the existing statistics (sampling methods) efficiently and thus is complementary to their approaches.

It is worth mentioning that, Gjoka et al. [19] designed a multi-graph sampling technique for the social networks which have multiple relation graphs. Their technique improves the convergence rate of the sampler by walking along a union graph of all relations. But it does not distinguish the efficiencies of walking on different relation graphs. In this paper, we propose the two-stage framework to select an inferred most efficient one from multiple graphs to improve sampling efficiency further.

## VII. CONCLUSIONS

In this paper, we consider the problem of using multiple statistics to efficiently sample online social networks. Given a fixed sampling budget, we design budget allocation decisions and combine them to construct an optimal estimator. In particular, we formulate a mixture sampling problem which constructs the optimal mixture estimator, and derive the optimal weights and a condition of ranking budget allocation decisions for the optimal estimator. Because the asymptotic variances of the individual statistics are unknown in practice, we propose an adaptive two-stage framework, which spends a partial budget to test all different statistics in the pilot sampling stage and allocates the remaining budget to the inferred best statistic in the regular sampling stage. To optimally set the sub-budget

for the pilot sampling stage, we design an adaptive algorithm to dynamically decide an upper-bound of the optimal pilot budget and test whether the pilot sampling should end. We implement the adaptive two-stage framework and evaluate its performance in the Douban network. We demonstrate, in theory and experiment, that our two-stage framework achieves higher sampling efficiency than two benchmark strategies.

## REFERENCES

- [1] A. H. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Respondent-driven sampling for characterizing unstructured overlays," *Proceedings of IEEE INFOCOM*, pp. 2701–2705, 2009.
- [2] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Practical recommendations on crawling online social networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 9, pp. 1872–1892, 2011.
- [3] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pp. 390–403, 2010.
- [4] K. Avrachenkov, B. Ribeiro, and D. Towsley, "Improving random walk estimation accuracy with uniform restarts," *Algorithms and Models for the Web-Graph*, pp. 98–109, 2010.
- [5] A. Mira, "Ordering and improving the performance of monte carlo markov chains," *Statistical Science*, pp. 340–350, 2001.
- [6] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 29–42, 2007.
- [7] Mohaisen, Abdelaziz and Yun, Aaram and Kim, Yongdae, "Measuring the mixing time of social graphs," *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pp. 383–389, 2010.
- [8] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, 2007.
- [9] Heckathorn, Douglas D, "Respondent-driven sampling: a new approach to the study of hidden populations," *Social problems*, pp. 174–199, 1997.
- [10] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore, "On the bias of traceroute sampling: or, power-law degree distributions in regular graphs," *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pp. 694–703, 2005.
- [11] L. Becchetti, C. Castillo, D. Donato, A. Fazzzone, and I. Rome, "A comparison of sampling techniques for web graph characterization," *Proceedings of the Workshop on Link Analysis*, 2006.
- [12] Gjoka, Minas and Kurant, Maciej and Butts, Carter T and Markopoulou, Athina, "Walking in Facebook: A case study of unbiased sampling of OSNs," *Proceedings of IEEE INFOCOM*, pp. 1–9, 2010.
- [13] M. Kurant, A. Markopoulou, and P. Thiran, "On the bias of BFS (Breadth First Search)," *22nd International Teletraffic Congress*, pp. 1–8, 2010.
- [14] Hastings, W Keith, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.
- [15] Henzinger, Monika R and Heydon, Allan and Mitzenmacher, Michael and Najork, Marc, "On near-uniform URL sampling," *Computer Networks*, vol. 33, no. 1, pp. 295–308, 2000.
- [16] Rasti, Amir H and Torkjazi, Mojtaba and Rejaie, Reza and Stutzbach, D, "Evaluating sampling techniques for large dynamic graphs," *Univ. Oregon, Tech. Rep. CIS-TR-08*, vol. 1, 2008.
- [17] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou, "Walking on a Graph with a Magnifying Glass: Stratified Sampling via Weighted Random Walks," *Proceedings of ACM SIGMETRICS*, 2011.
- [18] C.-H. Lee, X. Xu, and D. Y. Eun, "Beyond random walk and metropolis-hastings samplers: why you should not backtrack for unbiased graph sampling," *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 1, pp. 319–330, 2012.
- [19] M. Gjoka, C. T. Butts, M. Kurant, and A. Markopoulou, "Multigraph sampling of online social networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 9, pp. 1893–1905, 2011.
- [20] E. C. Titchmarsh, *The theory of functions*. London, 1939, vol. 80.

## APPENDIX

**Proof of Theorem 1:** From Equation (4), we have

$$\begin{aligned}\lim_{M \rightarrow \infty} \hat{f}(\mathbf{a}, M, \mathbf{w}) &= \lim_{M \rightarrow \infty} \sum_{k \in \mathcal{K}_a} w_k \cdot \hat{f}_k(a_k M) \\ &= \sum_{k \in \mathcal{K}_a} w_k \cdot \lim_{M \rightarrow \infty} \hat{f}_k(a_k M) \xrightarrow{a.s.} \sum_{k \in \mathcal{K}_a} w_k \bar{f}\end{aligned}$$

implying that the mixture estimator  $\hat{f}(\mathbf{a}, M, \mathbf{w})$  is asymptotically unbiased for  $\bar{f}$  if and only if  $\sum_{k \in \mathcal{K}_a} w_k = 1$ , i.e., Equation (8) concludes. Then from Equation (5), observe that

$$\begin{aligned}\varsigma(\mathbf{a}, \mathbf{w}) &= \lim_{M \rightarrow \infty} M \cdot \text{Var}(\hat{f}(\mathbf{a}, M, \mathbf{w})) \\ &= \lim_{M \rightarrow \infty} M \sum_{k \in \mathcal{K}_a} w_k^2 \cdot \text{Var}(\hat{f}_k(a_k M)) \\ &= \sum_{k \in \mathcal{K}_a} \frac{w_k^2}{a_k} \lim_{M \rightarrow \infty} a_k M \cdot \text{Var}(\hat{f}_k(a_k M)) = \sum_{k \in \mathcal{K}_a} \frac{w_k^2}{a_k} \cdot \sigma_k^2.\end{aligned}$$

Based on Cauchy-Schwarz inequality, it satisfies

$$\left[ \sum_{k \in \mathcal{K}_a} w_k^2 \cdot \frac{\sigma_k^2}{a_k} \right] \cdot \left[ \sum_{k \in \mathcal{K}_a} \frac{a_k}{\sigma_k^2} \right] \geq \left[ \sum_{k \in \mathcal{K}_a} w_k \right]^2 = 1$$

where the equality holds up if and only if  $w_k = \frac{a_k}{\sigma_k^2} / \sum_{i \in \mathcal{K}_a} \frac{a_i}{\sigma_i^2}$ . Thus given an allocation decision  $\mathbf{a}$ , for any weight vector  $\mathbf{w} \in \mathcal{W}_a$ ,

$$\varsigma_a(\mathbf{w}) = \sum_{k \in \mathcal{K}_a} w_k^2 \cdot \frac{\sigma_k^2}{a_k} \geq \left[ \sum_{k \in \mathcal{K}_a} \frac{a_k}{\sigma_k^2} \right]^{-1} = \varsigma_a(\mathbf{w}^*)$$

holds up, i.e.,  $\mathbf{w}^*$  solve the optimization problem in Equation (6).

**Proof of Theorem 2:** If the allocation decisions  $\mathbf{a}$  and  $\mathbf{a}'$  satisfies  $\sum_{k=1}^i a_{(k)} \geq \sum_{k=1}^i a'_{(k)}$  ( $i=1, \dots, K$ ),

$$\begin{aligned}\sum_{k=1}^K \frac{a_{(k)} - a'_{(k)}}{\sigma_{(k)}^2} &\geq \frac{a_{(1)} + a_{(2)} - a'_{(1)} - a'_{(2)}}{\sigma_{(2)}^2} + \sum_{k=3}^K \frac{a_{(k)} - a'_{(k)}}{\sigma_{(k)}^2} \\ &\geq \dots \geq \frac{\sum_{k=1}^K [a_{(k)} - a'_{(k)}]}{\sigma_{(K)}^2} = 0.\end{aligned}$$

holds up. Based on Theorem 1, we have

$$\varsigma(\mathbf{a}, \mathbf{w}^*(\mathbf{a})) = \left[ \sum_{k \in \mathcal{K}_a} \frac{a_{(k)}}{\sigma_{(k)}^2} \right]^{-1} \leq \left[ \sum_{k \in \mathcal{K}_a} \frac{a'_{(k)}}{\sigma_{(k)}^2} \right]^{-1} = \varsigma(\mathbf{a}', \mathbf{w}^*(\mathbf{a}')).$$

In particular, for any  $\mathbf{a}$ , the allocation  $\mathbf{a}^*$  satisfies  $\varsigma(\mathbf{a}, \mathbf{w}^*(\mathbf{a})) \geq \varsigma(\mathbf{a}^*, \mathbf{w}^*(\mathbf{a}^*)) = \sigma_{k^*}^2$ .

**Proof of Theorem 3:** When  $c = 1$ ,  $a_k(1) = 1/K$  ( $k = 1, \dots, K$ ). Then we have  $\varsigma(\mathbf{a}(1), \mathbf{a}(1)) = \frac{1}{K} \sum_{k=1}^K \sigma_k^2$  from Equation (9). When  $c = 0$ , the inferred most efficient statistic is uniform randomly chosen, i.e.,  $P(\hat{k}^*(cM) = k) = 1/K$  ( $k = 1, 2, \dots, K$ ). Then  $P(a_k(0) = 1) = 1/K$  and  $\varsigma(\mathbf{a}(0), \mathbf{w}^*(0)) = \sum_{k=1}^K P(a_k(0) = 1) \cdot \sigma_k^2 = \frac{1}{K} \sum_{k=1}^K \sigma_k^2$ .

**Proof of Theorem 4:** As each estimated asymptotic variance  $\hat{\sigma}_k^2(\cdot)$  is an asymptotically unbiased for  $\sigma_k^2$  ( $k = 1, \dots, K$ ), observe that

$$\begin{aligned}\lim_{M \rightarrow \infty} P(\hat{k}^* = k^*) \\ &= \lim_{M \rightarrow \infty} P(\hat{\sigma}_{k^*}^2(\frac{c(M)M}{K}) \leq \hat{\sigma}_j^2(\frac{c(M)M}{K}) \quad \forall j = 1, \dots, K) \\ &= \lim_{M \rightarrow \infty} P(\sigma_{k^*}^2 \leq \sigma_j^2 \quad \forall j = 1, \dots, K) = 1\end{aligned}$$

holds up if  $c(M) \in \omega(M^{-1})$ . Thus, we have, as  $M \rightarrow \infty$ ,

$$a_k(c(M)) \xrightarrow{a.s.} \frac{c(M)}{K} + (1 - c(M)) \cdot \mathbf{1}_{\{k=k^*\}} = a_k^*(M)$$

for  $\forall k = 1, \dots, K$ . Consequently, it satisfies  $\hat{k}^* \xrightarrow{a.s.} k^*$ ,  $\mathbf{a}(c(M)) \xrightarrow{a.s.} \mathbf{a}^*(M)$  and  $\hat{\mathbf{w}}(c(M)) \xrightarrow{a.s.} \mathbf{w}^*(\mathbf{a}^*(M))$  as  $M \rightarrow +\infty$ .

**Proof of Corollary 1:** From Theorem 1, for any  $c(M) \in \omega(M^{-1})$ , we have  $\varsigma(\mathbf{a}^*(M), \mathbf{w}^*(\mathbf{a}^*(M))) \leq \varsigma(\mathbf{a}^*(M), \mathbf{a}^*(M))$  as  $M \rightarrow \infty$ , where

$$\begin{aligned}\varsigma(\mathbf{a}^*(M), \mathbf{w}^*(\mathbf{a}^*(M))) &= \left[ \sum_{k=1}^K \frac{a_k^*(M)}{\sigma_k^2} \right]^{-1} \\ &\leq \left[ \frac{a_{k^*}^*(M)}{\sigma_{k^*}^2} \right]^{-1} = \frac{K \sigma_{k^*}^2}{K + (1 - K)c(M)}, \\ \varsigma(\mathbf{a}^*(M), \mathbf{a}^*(M)) &= (1 - c(M)) \sigma_{k^*}^2 + \frac{c(M)}{K} \\ &= \frac{1}{K} \sum_{k=1}^K \sigma_k^2 - \frac{1 - c(M)}{K} \sum_{k=1}^K (\sigma_k^2 - \sigma_{k^*}^2) \leq \frac{1}{K} \sum_{k=1}^K \sigma_k^2.\end{aligned}$$

**Proof of Theorem 5:** Because  $\mathbb{E}[\varsigma(\mathbf{a}(c(M)), \hat{\mathbf{w}}^*(c(M)))] = \lim_{M \rightarrow +\infty} M \cdot \text{Var}(\hat{f}(\mathbf{a}(c), M, \hat{\mathbf{w}}^*(c)))$  from Equation (5), the fraction  $c^*(M)$  minimizes  $\mathbb{E}[\varsigma(\mathbf{a}(c(M)), \hat{\mathbf{w}}^*(c(M)))]$ . When the convergence rate of estimated asymptotic variance  $\hat{\sigma}_k^2(m)$  for  $\sigma_k^2$  is  $\Theta(m^{-\eta_k})$  ( $k = 1, \dots, K$ ) and  $c(M) \in \omega(M^{-1})$ , it satisfies  $\mathbb{E}[\varsigma(\mathbf{a}(c(M)), \hat{\mathbf{w}}^*(c(M)))] \rightarrow \mathbb{E}[\varsigma(\mathbf{a}^*(M), \mathbf{w}^*(\mathbf{a}^*(M)))]$  as  $M \rightarrow +\infty$  with the convergence rate  $\Theta((c(M)M)^{-\eta})$  from Theorem 4 and Bounded Convergence Theorem [20]. Further, as  $c^*(M)$  minimizes  $\mathbb{E}[\varsigma(\mathbf{a}(c(M)), \hat{\mathbf{w}}^*(c(M)))]$ , it satisfies the first-order condition  $\lim_{M \rightarrow +\infty} \frac{d\mathbb{E}[\varsigma(\mathbf{a}(c(M)), \hat{\mathbf{w}}^*(c(M)))]}{dc(M)} \Big|_{c=c^*(M)} = 0$ , and therefore  $\Theta(\frac{d(cM)^{-\eta}}{dc} \Big|_{c=c^*(M)}) = \Theta(-\eta M^{-\eta} c^*(M)^{-\eta-1}) = \Theta(\frac{d\mathbb{E}[\varsigma(\mathbf{a}^*(M), \mathbf{w}^*(\mathbf{a}^*(M)))]}{dc} \Big|_{c=c^*(M)}) = \Theta(1)$ , from which we can derive that  $c^*(M) = \Theta(M^{\frac{-\eta}{\eta+1}}) = \Theta(M^{-1+\frac{1}{\eta+1}})$  and  $\lim_{M \rightarrow +\infty} c^*(M) = 0$ .